KSH Államigazgatási Számítógépes Szolgálat

Információelméleti módszeren alapuló lényegkiemelési eljárás és klinikai adatbázis redukciója

Balogh Gábor, Götl Győző és Srajber Benedek

Bevezetés

Az orvostudomány és a pszichológia gyakran használt vizsgálati eszközei a tesztek, amelyekkel a vizsgált tényezők vagy tényezőcsoportok egymásrahatásának mértékére, illetve fontosságuk sorrendjére vonhatunk le következtetéseket. Mivel azonban a tesztek kiértékelése során fellépő valószinüségi változók értékkészletei (azaz a kérdésekre adható válaszok) rendezetlen, diszkrét halmazok, ezért a statisztika klasszikus módszerei (korrelációszámitás, regresszióanalizis stb.) csak igen korlátozottan alkalmazhatók. A rendezetlen halmazokat rendezettekre leképezve ugyanis korrelációs vizsgálatokkal csak a lineáris függés erősségét mérhetjük. (Az eredmény természetesen a leképezéstől is függ.)

E problémák megoldására születtek meg a statisztika ujabb fejezetei és az alakfelismerés statisztikus módszerei: lényegkiemelés, hierarchikus- és szekvenciális Cluster-analizis stb.

Az általunk használt információs mérték diszkrét valószinüségi változók egymástól való függése tényleges nagyságának meghatározására szolgál. Mivel ennek kiszámitásához csak a változók együttes eloszlását leiró un. kontingencia-táblázat szükséges, emiatt attól is független, hogy a változókat nominális-, ordinális-, intervallum- vagy arányskálával jellemezzük. Az információs mérték képezi az alapját adatbázis-redukciós algoritmusunknak is.

1. Információs mérték

Legyen Z tetszőleges diszkrét, véges értékkészletü

$$p_i = p\,(Z = z_i) \qquad /i = 1,\ldots,n/ \text{ eloszlásu}$$

$$\sum_{i=1}^{n} p_i = 1$$

valószinüségi változó. Ekkor Z kimenetelének bizonytalanságát entrópiája, azaz a

$$H(Z) = \sum_{i=1}^{n} p_i \log \frac{1}{p_i}$$ mennyiség méri.

Ha Z vektorváltozó, azaz Z = (X,Y), ahol X és Y valószinüségi változók, akkor (a Jensen-egyenlőtlenség alkalmazásával) a

$$H(Z) = H(X,Y) \leq H(X) + H(Y)$$

egyenlőtlenséghez jutunk. Az egyenlőség csak X és Y függetlensége esetén teljesül /lásd (3)/.

Az egyenlőtlenség két oldala közti különbséget kölcsönös információnak nevezzük és I(X,Y) -nal jelöljük. Az elnevezést az indokolja, hogy az X változó éppen I(X,Y) bit információt tartalmaz az Y-ra vonatkozóan (és ugyanennyit tartalmaz Y az X-re nézve).

Tehát az $I(X,Y) = H(X) + H(Y) - H(X,Y)$ bevezetésével az $I(X,Y) \geq 0$. Ismeretes, hogy $I(X,Y) \leq H(Y)$ ill. $I(X,Y) \leq H(X)$ is fennáll, továbbá pl. az $I(X,Y) \leq H(X)$ -ben pontosan akkor van egyenlőség, ha X az Y függvénye.

Következésképpen az $i(X,Y) = \frac{I(X,Y)}{H(X)}$ formulával meghatározott mennyiségre $0 \leq i(X,Y) \leq 1$, $i(X,Y) = 1$ akkor és csak akkor, ha X függvénye Y-nak. Igy i(X,Y) olyan nem-szimmetrikus mennyiség, mely X-nek Y-tól való függését méri, származtatása alapján <u>információs mértéknek</u> nevezzük.

Természetesen, két valószinüségi változó kapcsolatának részletesebb vizsgálatánál nem lehet elegendő csupán az információs mérték és a korrelációs értékének meghatározása. Ezek ugyanis csak a két változó <u>globális</u> kapcsolatát jellemzik. Azt, hogy az X változó <u>egy konkrét</u> kimenetele milyen mértékben határozza meg Y-t, az $i(Y,X=x_i)$ <u>parciális információs mérték mutatja meg:</u>

$$i(Y,X=x_i) = \frac{H(Y) - H(Y,X=x_i)}{H(Y)}$$

Könnyen belátható, hogy

$$i(Y,X)=\sum_{j=1}^{n} p(X=x_j) \cdot i(Y,X=x_j) \text{ és } 1 - \frac{\log k}{H(Y)} \leq i(Y,X=x_j) \quad 1 .$$

Ha $i(Y,X=x_j) = 1$, akkor az $X=x_j$ kimenetel teljes mértékben meghatározza Y kimenetelét, ha $i(Y,X=x_j) = 0$, akkor bizonytalanságunk változatlan marad, ha pedig negativ, akkor a szóban forgó megfigyelés növeli bizonytalanságunkat.

Legyenek most X_j-k $/j = 1,\ldots,k/$ és Y valószinüségi változók. Igazolható, hogy

$$I(Y,X_1,\ldots,X_k) - I(Y,X_1,\ldots,X_{k-1}) = \\ = I(X_k,X_1,\ldots,X_{k-1},Y) - I(X_k,X_1,\ldots,X_{k-1}), \qquad /1/$$

amely azt mutatja, hogy X_k kihagyásával Y-ról annyi információt vesztünk, amennyit X_k-ról nyerhetünk Y megismerése által. Az előző egyenlőség segitségével kapható az

$$i(Y,X_1,X_2) \geq i(Y,X_1) + i(Y,X_2) - \frac{I(X_1,X_2)}{H(Y)}$$

egyenlőtlenség. Egyenlőség akkor és csak akkor van, ha

$$i(X_2,Y,X_1) = i(X_2,Y),$$

ami azzal ekvivalens, hogy "X_1 és X_2 független Y bármely kimenete esetén". (Megjegyezzük, hogy ennek semmi kapcsolata a valószinüségi változók között értelmezett függetlenség fogalmával, amint azt Lee (2) munkájában feltételezte.) Igy független valószinüségi változók esetén is csak az $i(Y,X_1,X_2) \geq i(Y,X_1) + i(Y,X_2)$ egyenlőtlenség igaz.

Az információs mérték közelitésével és a közelités szignifikanciájának vizsgálatával az (1) tanulmány foglalkozik. Ezeket a problémákat a következőképpen foglalhatjuk össze:

Jelöljük két valószinüségi változó információs mértékének elméleti értékét i(Y,X)-szel, gyakorlati meghatározására N számu kisérletet végzünk. Azaz N elemü mintapopuláció alapján meghatározzuk az (Y,X) tapasztalati eloszlásfüggvényét és ennek segitségével kiszámitjuk az ehhez a mintakollekcióhoz tartozó tapasztalati információs mértéket,

$$i^{N}_{tap}(Y,X) \text{ -et.}$$

Belátható, hogy a mintaanyag nagyságának növekedésével a tapasztalati információs mérték sztochasztikusan konvergál az információs mérték elméleti értékéhez.

II. Adatredukciós algoritmusunk elméleti meggondolása

Az előző fejezetben megtárgyaltuk az információs mérték tulajdonságait és utaltunk becslésének módjára is. Ebben a részben rámutattunk az információs mérték fontosságára a lényegkiemelő és adatredukciós módszerek egy osztályában.

Az Y és az X, véges számu diszkrét értéket felvevő valószinüségi változók függetlenségét vizsgáljuk.

Null-hipotézisünk /H_o/ : Y és X függetlenek,

ellen-hipotézisünk /H_e/: nem-függetlenek

$\{p(Y_i) \cdot p(x_j)\}$ illetve $\{p(Y_i, x_j)\}$ eloszlással.

Döntés céljából n számu független megfigyelést végzünk. Egy megfigyeléssorozat eredményét az

$$u = (Y_{i_1}, x_{i_1}, Y_{i_2}, x_{i_2}, \ldots, Y_{i_n}, x_{i_n})$$

jellemzi. A megfigyelések függetlensége miatt természetesen:

$$p^n(u \mid H_o) = \prod_{k=1}^{n} p(Y_{i_k}) \cdot p(x_{i_k}) \text{ ill. } p^n(u \mid H_e) = \prod_{k=1}^{n} p(Y_{i_k}, x_{i_k}) .$$

Jelöljük az n megfigyelés összes lehetséges kimeneteleinek halmazát R-rel és legyen E (elfogadási tartomány) R-nek egy részhalmaza.

Induljunk ki abból az elvből, hogy $u \in E$ esetén H_o, $u \notin E$ esetén pedig H_e javára döntünk. Célunk akkor az E halmaz olyan meghatározása, hogy a helytelen döntésből eredő hiba minél kisebb legyen. Tehát olyan E halmazt keresünk, melynél az elsőfaju

$$p^n(E^c|H_o) = \sum_{u|u \notin E} p^n(u|H_o) ,$$

ill. a másodfaju

$$p^n(E|H_e) = \sum_{u|u \in E} p^n(u|H_e)$$

hibák minimálisak. Mivel ezek változása ellentétes irányultságu, azért minimumuk egyszerre nem valósitható meg. Ehelyett - rögzitett esetén - vizsgálhatjuk a következő kifejezés n-től való függését:

$$\gamma(n) = \inf_{\substack{E: E \subset R \text{ és} \\ p^n(E|H_o) \geq 1-\beta}} p^n(E|H_e) ,$$

azaz azt, hogy az elsőfaju hiba valószinüségét szignifikancia-szinten tartva hogyan függ a minimális másodfaju hiba valószinüsége a mintaanyag nagyságától. Erre vonatkozik a Chernofftól származó tétel (6):

$$\lim_{n\to\infty} \frac{1}{n} \cdot \log \gamma(n) = - I(Y,X) .$$

Tehát, az elsőfaju hiba valószinüségét szignifikancia-szinten tartva a másodfaju hiba valószinüsége a mintaszám növelésével exponenciálisan csökkenthető. A csökkenés mértékét az I(Y,X) kölcsönös információ határozza meg. Igy, ha X vektorváltozó, azaz $X = (X_1,\ldots,X_s)$, akkor adatredukálás céljából azon $X_{r_1},\ldots,X_{r_m}$ változók kiválasztása kivánatos, melyekre az $I(Y,X_{r_1},\ldots,X_{r_m})$ kölcsönös információ és ebből következően az $i(Y,X_{r_1},\ldots,X_{r_m})$ információs mérték maximális.

III. Adatredukciós algoritmus

A feladat azon minimális számu $X_{r_1},\ldots,X_{r_m}$ változóknak az $X_1,\ldots,X_s$ változók közül történő kiválasztása, amelyek Y-ról az összes változó által nyerhető információnak "elég nagy részét" tartalmazzák. Pontosabban: keressük azokat az $X_{r_1},\ldots,X_{r_m}$ változókat, amelyekre

$$\left\{X_{r_j}\right\}_{j=1}^{m} \subset \left\{X_i\right\}_{i=1}^{s} \qquad \text{és}$$

$i(Y,X_{r_1},\ldots,X_{r_m}) \geq (1-\varepsilon)\cdot i(Y,X_1,\ldots,X_s)$ feltételek mellett m értéke minimális.

A feladat egzakt megoldhatóságának két fő akadálya van:

1.) $i(Y,X_1,\ldots,X_s)$ gyakorlatilag kiszámithatatlan, értékét csak becsülni lehet.

2.) Az összes lehetséges változó-kombinációk leszámlálása gyakorlatilag elvégezhetetlen.

Ezen okok miatt nem is várható, hogy létezik optimális megoldáshoz vezető, gyors algoritmus. A következőkben leirt "ad hoc" módszer célja ezért nem az optimális, hanem egy ahhoz nagy valószinüséggel közel eső megoldás keresése.

Iterációs algoritmusunk lépései:

1.) Kiválasztjuk az $i(Y,X_{r_1}) = \max_{1\leq j\leq s} i(Y,X_j)$ feltételnek eleget tevő X_{r_1}-et.

2.) Meghatározzuk az $i(Y,X_{r_1},X_{r_2}) = \max_{\substack{1\leq j\leq s \\ j\neq r_1}} i(Y,X_{r_1},X_j)$ feltételt kielégitő X_{r_2}-t.

⋮

k.) Az előző lépésekben meghatározott $X_{r_1},\ldots,X_{r_{k-1}}$ változókat rögzitve

meghatározzuk azt az ujabb X_{r_k} változót, amelyre

$$i(Y,X_{r_1},\ldots,X_{r_k}) = \max_{\substack{1\leq j\leq s \\ j\neq r_q \ /q=1,\ldots,k-1/}} i(Y,X_{r_1},\ldots,X_{r_{k-1}},X_j) \qquad /2/$$

teljesül. Természetesen az információs mérték értékét az I. részben leirtak alapján tapasztalati értékével közelitjük.

A befejezhetőség kritériumai a következők:

a.) A kiválasztott változók által Y-ról nyert információ eléggé megközeliti-e az összes kapható információt?

b.) Feltételezhető-e, hogy ujabb változó kiválasztásával még további szignifikáns mennyiségü információt nyerünk?

Az algoritmus számitógépes realizálásával kapcsolatban szükségesnek tartjuk megjegyezni a következőket:

a.) A lépésenkénti számolás mennyisége a lépésszám növekedésével exponenciálisan nő, ezért a kiválasztható változók száma elsősorban a mintaanyag nagyságától és a feldolgozó számitógép paramétereitől függ.

b.) Az /1/ összefüggés figyelembevétele alapján az algoritmussal kiválasztott változók egymástól csak kevéssé függhetnek, mivel módszerünk csak akkor választ ki egy, az előzőktől erősen függő változót, ha másképpen nem kaphatnánk Y-ról legalább ugyanannyi információt.

c.) A megoldhatóság nehézségei miatt a befejezést eldöntő vizsgálatokat függvény approximációs módszerekkel végezhetjük el.

d.) Az algoritmus részletes leirása és FORTRAN programja az ÁSZSZ Honeywell 66/60-as gépének programkönyvtárában az érdeklődők rendelkezésére áll. A program próbafeladaton sikeresen lefutott. Nagyméretü mintaanyagon kipróbálása folyamatban van.

IV. Klinikai adatbázis redukciója, differenciál-diagnózis és prognózis

A diagnosztikai folyamat végrehajtása közben az orvos rendszerint szembetalálja magát azzal a nehézséggel, hogy a nagy számosságu evidencia halmaz nehezen kezelhető vagy egyszerüen áttekinthetetlen számára. Ebből a tényből hármas probléma adódik:

a.) Az evidencia halmaz elemeinek feltétlen szükséges redukciója

Ezt megoldhatjuk oly módon, hogy a III. pontban ismertetett algoritmus /2/ képletébe Y-nak az $Y = (X_1,\ldots,X_s)$ valószinüségi vektorváltozót választjuk, ahol X_j-k $(j=1,\ldots,s)$ mint valószinüségi változók, a beteg adatainak felelnek meg. Az eredmény: a betegre vonatkozó tényezők számának csökkenése oly módon, hogy a redukált tényezők halmaza közel azonos információ-mennyiséget tartalmaz, mint a kiindulásul vett tényezők együttesen.

b.) A differenciál-diagnózis megállapitásához szükséges legszükebb evidencia halmaz megállapitása

Ha egy klinikai adatbázis esetén a /2/ képletben Y a szóbajövő betegség-osztályokat reprezentálja, akkor az ismertetett adatredukciós eljárás az összes vizsgált tényező közül azon minimális számu "közel független" tényező kiválasztását eredményezi, amely elegendő információt tartalmaz a betegség-osztályok valószinüsitéséhez. Kevés számu tényező esetén most már a betegek betegség-osztályba sorolásához a jól ismert Bayes-féle differenciál-diagnosztikai modellel (9) vagy más módszerekkel is könnyebben eljuthatunk.

c.) A beteg prognózisához (pl. tulélések valószinüsitése) szükséges legszükebb halmaz meghatározása

Ha a /2/ képletben Y - mint valószinüségi változó - valamely betegség összes lehetséges kimeneteleit, X_j-k $(j=1,\ldots,s)$ pedig a kimenetelt befolyásoló összes szóbajöhető tényezőt reprezentálják, az eredmény b.)-hez (adatredukció és osztályba sorolás) hasonlóan a kivánt redukált halmaz és a megfelelő kimeneteli kategóriába való besorolás.

Irodalom

(1) Andor Cs., Mérő L.: Kölcsönös információ alkalmazása két vagy több változó kapcsolatának meghatározására. MRT Tud közl.

(2) R. Char-Tung Lee: Application of Information Theory to Relevant Variables. Math. Biosci. 11, 153-161 (1971)

(3) Csiszár I., Fritz J.: Információelmélet, 1970.

(4) Fritz J.: Bevezetés az információelméletbe (Tankönyvkiadó)

(5) J. Kryspin, A. M. Norwich: Use of Information Theory in Analysis of Medical Data. 1972 IEEE Conference on Inf. Theory, Asilomar, California.

(6) S. Kullback: Information Theory and Statistics. Whiley, 1959.

(7) W. S. Meisel: Computer-oriented Approaches to Pattern Recognation. Acad. Press, New York, 1972.

(8) J. M. Mendel, K. S. Fu: (szerk) Adaptive, Learning and Pattern Recognition System. Acad. Press, 1970.

(9) L. G. Withby, W. Lutz: Principles and Practice of Medical Computing.